MIS Quarterly

RESEARCH ARTICLE

CONTENT SHARING IN A SOCIAL BROADCASTING ENVIRONMENT: EVIDENCE FROM TWITTER[1]

Zhan Shi
Department of Information Systems, W. P. Carey School of Business, Arizona State University,
Tempe, AZ 85287 U.S.A. {zhan.m.shi@asu.edu}

Huaxia Rui
Simon Graduate School of Business, University of Rochester, Rochester, NY 14627 U.S.A. {huaxia.rui@simon.rochester.edu}

Andrew B. Whinston
McCombs School of Business, The University of Texas at Austin, Austin, TX 78712 U.S.A. {abw@uts.cc.utexas.edu}

The rise of social broadcasting technologies has greatly facilitated open access to information worldwide, not only by powering decentralized information production and consumption, but also by expediting information diffusion through social interactions like content sharing. Voluntary information sharing by users in the context of Twitter, the predominant social broadcasting site, is studied by modeling both the technology and user behavior. A detailed data set about the official content-sharing function on Twitter, called ***retweet****, is collected and the statistical relationships between users' social network characteristics and their retweeting acts are documented. A two-stage consumption-sharing model is then estimated using the conditional maximum likelihood estimatio (MLE) method. The empirical results convincingly support our hypothesis that weak ties (in the form of unidirectional links) are more likely to engage in the social exchange process of content sharing. Specifically, we find that after a median quality tweet (as defined in the sample) is consumed, the likelihood that a unidirectional follower will retweet is 3.1 percentage point higher than the likelihood that a bidirectional follower will do so.*

Keywords: Content sharing, social broadcasting, information diffusion, Twitter, weak tie

Introduction

At 10:24 p.m. EST, May 1, 2011, one hour and eleven minutes before the formal announcement of Osama bin Laden's death by U.S. President Barack Obama, the following message was posted on Twitter by Mr.Keith Urbahn:[2]

> *So I'm told by a reputable person they have killed Osama Bin Laden...*

The post quickly attracted attention and was forwarded by Mr.Urbahn's subscribers on Twitter, and within two minutes, there were already more than 300 reactions to it. In the following hour, tens of thousands more users in the Twitter world were passing this message, and the final number of people who were exposed to the information *before* the formal White House announcement was even higher.

This example not only shows the sheer power of Twitter as a fast-growing *social medium*, but also demonstrates that the

[1] Ravi Bapna was the accepting senior editor for this paper. Ramesh Sankaranarayanan served as the associate editor.

The appendices for this paper are located in the "Online Supplements" section of the *MIS Quarterly*'s website (http://www.misq.org).

[2] @keithurbahn, http://twitter.com/keithurbahn.

emerging social media can beat even their mainstream competitors in terms of speed, flexibility, and reach, especially in tracking events as they unfold in real time.[3] The unique advantage of websites like Twitter in disseminating news comes from their distinctive technological infrastructure. Although Twitter and a number of other similar online services, such as Tumblr and Sina Weibo, are usually referred to as microblogging or social networking sites, these labels fail to capture their whole essence—that these websites each are simultaneously a broadcasting service and a social network. Like content from most traditional mass media, *user-generated content* on these sites is accessible by the public and is broadcast through directed subscription. The subscription relationships, as the only kind of user relationship, constitute the accompanying social network. The coexistence of a broadcasting service and a social network makes the combination of facets easily distinguishable from each one's respective standalone peers. On the one hand, the broadcasting service differs from traditional mass media like TV or radio in its decentralized structure and its social ingredient; it represents the full spectrum of communications, from headline news to personal and private communications (Wu et al. 2011). On the other hand, the social network, derived from content-subscription relationships, also significantly differs from traditional online social networks, which typically map real-world friendships or connections. For example, the social network on Twitter is quite open and loose compared to the social network on Facebook because the follower– following relationship on Twitter can be established unilaterally and usually cuts across long (real-world) social distances. This combination gives these technologies unique advantages in facilitating information diffusion and justifies assigning them to a new category, which we call *social broadcasting networks*. This view is also explicitly or implicitly shared by many computer and information scientists. For example, Kwak et al. (2010) suggested that Twitter more closely resembles an information sharing site than a traditional social network. Bakshy et al. (2011) noted that "unlike other user-declared networks…Twitter is expressly devoted to disseminating information" (p. 65). Social broadcasting networks have blurred the traditional boundary between social networks and news media by adding the "social" ingredient into the cycle of information production, exchange, and consumption (Kwak et al. 2010; Lotan 2011; Wu et al. 2011).

As exemplified by the bin Laden case, information diffusion in social broadcasting networks critically relies on social interactions, such as content sharing. Indeed, without the voluntary relaying of Mr. Urbahn's message by numerous Twitter users, that single post might never have triggered an avalanche of reactions and reached an audience far beyond Mr. Urbahn's own subscribers.[4] Content sharing is a critical mechanism of information diffusion in social broadcasting networks and is vital to a network's proper functioning and thriving. When interesting or important information does not get passed on, the social broadcasting network fails to reach its full potential as a news medium; meanwhile, excess transmission of redundant or trivial information creates information overload and lowers the value of a social broadcasting network to the users. Understanding the information relaying process is thus both interesting and important. The objective of this paper is to make an early step in this direction by examining the decision-making process of sharing at the individual level. As suggested, one defining feature of social broadcasting networks is that they possess a large volume of weak interpersonal relationships. Thus, our central goal in this article is to address the following research question:

> *How does the strength of the interpersonal tie moderate people's voluntary content sharing behavior in a social broadcasting network?*

Exploring the question might further reveal people's motivation in passing on information.

Users' voluntary content sharing is a social exchange process (Blau 1964) that involves the content's creator, the sharer, and the sharer's subscribers. To motivate our empirical exploration, we examine how tie strength moderates people's decisions to engage in the social exchange by drawing on two streams of prior research: the literature on tie strength and the literature on people's pro-social behavior.

Plenty of literature has looked at the implications of tie strength in a variety of social or economic settings. For example, Granovetter (1973) did the pioneering work on the role that weak ties play when people search for jobs, the result of which is famously summarized as *the strength of weak ties* (SWT). The arguments of SWT suggest the importance of weak ties (i.e., ties with acquaintances, rather than close friends) in enabling novel information to flow across two densely knit groups of close friends. Weenig and Midden (1991) studied the information diffusion process in two Dutch neighborhoods and their regression results seem to suggest

[3] Indeed, this capability has been proven again and again during events such as the 2009 Iran election, the 2011 Middle East Revolution, and the 2012 Chinese political scandal.

[4] According Lotan of the social media company SocialFlow, Keith Urbahn was not the first to speculate bin Laden's death after the news was released about the presidential address. However, Urbahn's tweet proved to be a watershed in people's discussion on Twitter regarding the presidential address.

that even in small communities new information that originates from outside the community diffuses in a community through weak ties rather than through strong ties. They could not distinguish whether this is due to the bridging capacity or the sheer number of weak ties because, unlike our study, they did not observe the actual path of information flow. Levin and Cross (2004) proposed and tested a model of dyadic knowledge exchange taking into account trust and tie strength between the two parties. Their results also suggest that weak ties provide access to nonredundant information. Bapna et al. (2012) studied the link between strength of social ties and trust in an online social network using data from a Facebook application. They found that, for the average user, social tie strength as measured by actively interacting with someone else is positively linked to trust.

Researchers have also studied extensively people's motivation of sharing knowledge in an online environment where explicit financial compensation is often absent (Bock et al. 2005; Chiu et al. 2006; Olivera et al. 2008; Wasko and Faraj 2005). Most of the previous studies focus on sharing behavior in the form of helping others (often strangers) solve problems by contributing one's *own* knowledge. Bock et al. (2005) surveyed 154 managers from 27 Korean organizations and found that anticipated reciprocal relationships affect individual's attitudes toward knowledge sharing. Chiu et al. (2006) also found that social interaction ties, reciprocity, and identification increased individuals' quantity of knowledge sharing by surveying 310 members of one professional virtual community in Taiwan. Olivera et al. (2008) developed a framework for understanding contribution behaviors and delineated three mediating mechanisms: awareness, searching and matching, and formulation and delivery. The sharing behavior we study is people's voluntary information relaying decision, which is a quite different type of contribution. Wasko and Faraj (2005) applied theories of collective action to examine how individual motivations and social capital influence knowledge contribution in electronic networks. Using survey data and archival data from one electronic network supporting a professional legal association, they found that people contribute their knowledge when they perceive that it enhances their professional reputations, when they have the experience to share, and when they are structurally embedded in the network. The current paper can be viewed as an extension of Wasko and Faraj in the sense that we are also examining people's contribution behavior on a electronic network. However, this paper departs from previous IS literature in two important ways. In terms of data and method, we use micro-level data and a two-stage discrete choice model to study a relatively new form of sharing behavior—relaying information contributed by others—in a social broadcasting network that is also a new form of virtual community. In terms of theoretical motivation, we integrate SWT with the general framework of social exchange to investigate the particular role of the tie strength in moderating the sharing behavior.

Our theoretical discussion posits the idea that one's motivation for engaging in the social exchange process of content sharing is the latent benefit of perceived reputation enhancement resulting from consumption of the shared content by one's subscribers. The perceived reputation enhancement is positively associated with the perceived novelty of the content to the sharer's subscribers, which in turn is negatively associated with the strength of the social tie between the content creator and sharer. The resulting hypothesis is that weak-tie subscribers are more likely to engage in sharing. Our hypothesis extends SWT in the social broadcasting context. The SWT theory argues that because weak ties are more likely to possess nonredundant information, information seekers (e.g., the job hunters in the classic example) should turn to weak ties for it. However, it is not clear whether weak ties are necessarily more likely to promote the flow of novel information in a proactive way. We quote the following paragraph from Friedkin (1980, p. 421):

> Granovetter's theory, to the extent that it is a powerful theory, rests on the assumption that local bridges and weak ties not only represent opportunities for the occurrence of cohesive phenomena... but that they actually do promote the occurrence of these phenomena. A major empirical effort in the field of social network analysis will be required to support this aspect of Granovetter's theoretical approach....It is one thing to argue that when information travels by means of these ties it is usually novel and, perhaps, important information to the groups concerned. It is another thing to argue that local bridges and weak ties promote the regular flow of novel and important information in differentiated structures. One may agree with the former and disagree with the latter.

To a certain extent, our work closes the gap between the two things Friedkin tried to disentangle conceptually, by arguing that social exchange motivates weak ties to facilitate the penetration of novel information in the context of social broadcasting networks.

Empirically testing our hypothesis in a real-world social broadcasting network is complicated both by the challenge of collecting micro-level data from the Internet and by the specifics of the actual technological environment in which data are generated. To overcome these problems, we deploy 20 computers over a 140-day period to collect a detailed data set containing information on both the content-sharing activity and social relationships from Twitter, and we develop a two-stage "consumption-sharing" model to help us better understand the machine-mediated human decision-making process. We then estimate the empirical model using the conditional maximum likelihood estimation (MLE) method, the results of which convincingly support our theory.

The remainder of this paper proceeds as follows. In the next section, we briefly introduce Twitter as an example of a social broadcasting network and describe the technology-mediated information-sharing mechanism on Twitter. Drawing on social and behavioral theories, we develop our hypothesis. After describing our data set, we conduct a series of empirical analyses to test our model, and we discuss the managerial implications of our findings. Finally, we conclude the paper and discuss future research directions.

Twitter and Retweeting

Designed to be the "Short Message Service of the Internet" at start-up, Twitter was launched in July 2006. During the 2007 South by Southwest (SXSW) Festival in Austin, TX, a showcase of Twitter impressed the highly tech-savvy attendees. Since then, Twitter has entered a phase of rapid growth and gained popularity far beyond the technology industry insiders. As of April 2012, Twitter had more than 500 million registered users worldwide, who in total posted an average of 340 million updates a day.[5] Twitter is now one of the most vibrant online communities in the world.

Twitter: A Social Broadcasting Technology

Twitter is an example of a social broadcasting site, where a broadcasting service and a social network organically constitute the technological infrastructure. On top of that, Twitter users produce and consume informational content by authoring and reading *tweets*,[6] which are text-based updates/messages of up to 140 characters. Like content on most traditional mass media, tweets are by default open to the public, and there is no restriction on consumption. Powered by its service, every Twitter user can be a content broadcaster and/or a content consumer.

Twitter users are networked to each other through a *following–follower* relationship. A user's *followers* are those who subscribe to receive his or her tweets, and a user's *followings* are the users whose tweets he or she subscribes to receive.[7] This following-follower relationship is the sole interpersonal link in the Twitter network. It is not only the pathway through which broadcasted content traverses the Twittersphere but also the channel of person-to-person communications, such as public reply and direct message. This relationship differs from *friendship* on Facebook or some other social network site in two respects: (1) the following–follower relationship on Twitter is relatively open in the sense that *A* following *B* does not require *B*'s consent, and they usually do not map to real-world friendships as the ones on Facebook do;[8] and (2) perhaps more importantly, the following–follower relationship is directed (*A*'s following *B* does not imply *B*'s following *A*) while friendship is undirected (*A*'s being a friend of *B* implies *B*'s being a friend of *A*). The existence of a large volume of loose and directed subscription relationships is thus a distinctive characteristic of a social broadcasting network.

Retweeting: Content Sharing on Twitter

Content sharing is an integral part of the Twitter experience. In addition to composing and posting tweets themselves, Twitter users can also rebroadcast—or *retweet*,[9] in Twitter's terminology—other users' (most likely their followings') tweets that they find are of particular (informational, entertaining, etc.) value.[10] Retweeting spreads information by exposing a new audience to the content. Meanwhile, retweeting is a special kind of sharing because a retweet is simply a copy of the original tweet, and thus the author, con-

[5] http://en.wikipedia.org/wiki/Twitter.

[6] Tweet can also be used as a verb, meaning to post. So "to tweet a tweet" means "to post an update."

[7] A user *A* does not have to follow *B* to consume *B*'s tweets. *A* can access *B*'s Twitter web page at any time to consume *B*'s tweets, which, like everyone else's, are always publicly available. But if *A* follows *B*, *B*'s tweets will be "pushed" to *A* in real time.

[8] The fact that users who are connected in a social broadcasting site are usually neither friends nor even acquaintances in the real world allows us to narrow our focus to just the online context in studying their interactions. For instance, we do not have to worry that a favor *A* does for *B* online would be reciprocated offline.

[9] *Retweet* is both a verb and a noun, just as *tweet* is. When user *A* retweets a tweet *t*, we call the reposted copy of *t* a *retweet* and call *A* a *retweeter* of *t*.

[10] Posting others' tweets simply by copying and pasting their tweets without mentioning the original author is technologically possible but is not considered retweeting. Rather, it is a highly criticized misbehavior in the Twitter community.

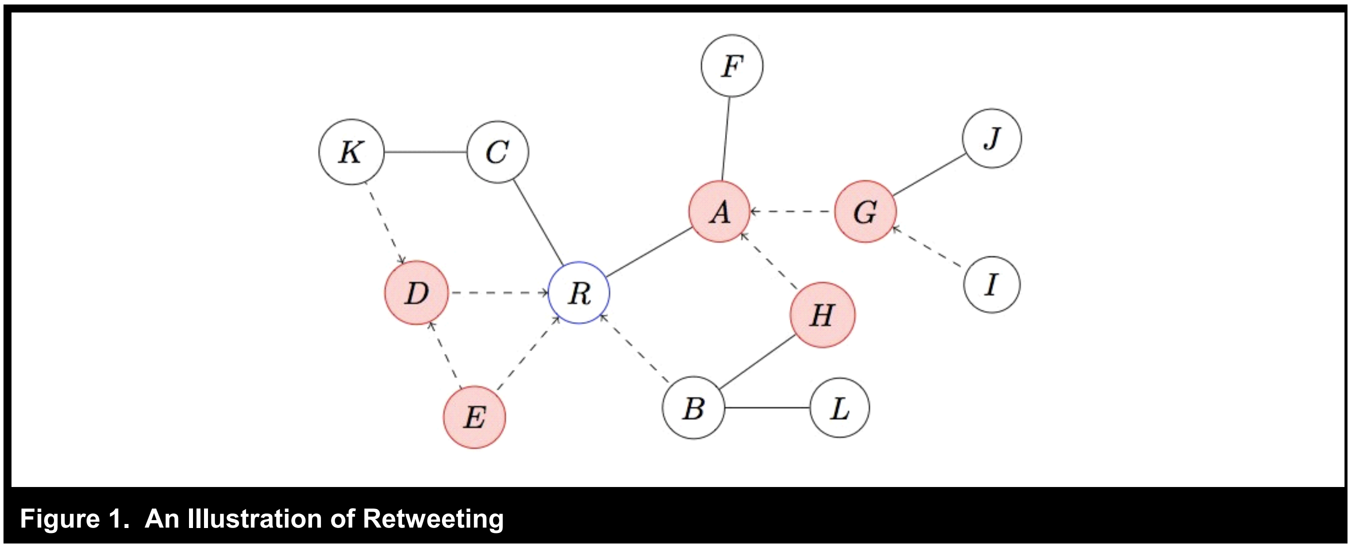


Figure 1. An Illustration of Retweeting

tent, and format of the shared information stay exactly the same as the original tweet. Retweeting can also display a "chain effect": not only a tweet's author's followers, but also sharers' followers, and so on, can further retweet, spreading the content onto their respective networks and amplifying the audience of the content to a potentially massive scale (Lotan 2011). Thus, retweeting is evidently a critical mechanism of information diffusion on Twitter. Since it was introduced, retweeting has been extremely popular on Twitter because of the straightforward idea and the easy-to-use official retweet button.[11] Therefore, we use retweeting in the Twittersphere as the primary real-world example of content sharing activity.[12]

The mechanism of retweeting is graphically illustrated in Figure 1. Hereafter, we call the user who writes the original tweet the *author*, and the *author* is denoted *R* in the figure. The other nodes represent other users who are linked to each other via the following–follower relationship, together forming a tiny community inside the Twitter world. If two users mutually follow each other, the line between them is solid (e.g., *R* and *A*, and we call *A* a bidirectional follower of *R*). Otherwise, if only one of them follows the other, the line between them is dashed, with an arrow pointing to the user followed (e.g., *B* follows *R* but *R* doesn't follow *B*, so that we call *B* a unidirectional follower of *R*). After *R* posts an update, if no one retweets it, only *R*'s followers *A*, *B*, *C*, *D*, and *E* would receive it. But now assume that after reading the message, users *A*, *D*, and *E* retweet (retweeters are designated by shaded circles), thereby making *F*, *G*, *H*, and *K*, who are not immediate followers of *R*, receive a copy of the tweet. Then the new receivers could also retweet (as *G* and *H* do in the Figure 1 example), circulating the information more broadly on the network. One thing to note is that a retweet is also a content broadcast; because of the technology, a sharer cannot select a subgroup of his or her followers and only retweet to this subgroup.[13]

Using the graphic example in Figure 1 as the context, we emphasize a few things related to our research question. First, we do not consider network dynamics (the formation and destruction of personal relationships among the users). In this research, we take a snapshot of the network structure, consider it as fixed and exogenous, and study user behavior on top of it. Second, in later econometric analyses, we model potential retweeters only in the first order (i.e., *R*'s immediate followers *A*, *B*, *C*, *D*, and *E*), but not those in the second and higher orders (i.e., *F*, *G*, *H*, *I*, *J*, and *K*). As we explain in the data section, the reason is that we do not have the network graph data for higher order potential sharers. Third, the variation of user behavior we exploit is different users' different reactions to a single tweet (e.g., the reactions of *A*, *B*,

[11] The official retweet function is built into most mobile applications, as well as Twitter's official website. There is no publicly available statistic on the popularity of retweeting versus other ways of sharing information. For example, another widely adopted way is to quote a tweet and add "RT" in front. An off-the-record interview with a Twitter employee confirmed that the official retweeting button had been the more popular mode of sharing.

[12] In addition to Twitter's dominance in the social broadcasting domain, another important reason we focus on it is that the openness of Twitter allows us to collect a detailed, micro-level data set to complete our study. The "Data" section describes our data collection in detail.

[13] In *non-broadcasting* social networks, such as Facebook, users typically can post messages only to a chosen subgroup of "friends."

C, *D*, and *E* to a tweet authored by *R*), rather than one single user's different reactions to different tweets (e.g., *B*'s reactions to different tweets authored by *R*, *H*, and *L*).

Theoretical Motivation

In this section, we develop the hypothesis about how the strength of interpersonal ties moderates people's decisions on relaying the messages of others. Although we often refer to Twitter as we develop our hypothesis, our theoretical arguments are applicable to other social broadcasting networks as well.

Content sharing is a social exchange process (Blau 1964; Homans 1958) that involves three parties: the sharer, the creator of the content, and the group of individuals with whom the content is shared. By choosing to relay the information, the sharer incurs the cost of sharing[14] without being rewarded in any explicit way. However, the other two parties explicitly benefit: Subscribers can consume the shared information, and the content creator reaches a larger audience.

Social exchange theory posits the idea that people engage in social exchange with the expectation of getting returns. When no explicit material or financial gains are received, the latent benefit of a social exchange process can be emotional comforts or social rewards such as approval, status, and respect (Wasko and Faraj 2005). Indeed, "people's positive sentiments toward and evaluations of others, such as affection, approval, and respect, are rewards worth a price that enter into exchange transactions" (Blau 1964, p. 112). Certain acts conducted by members of a community, such as sharing knowledge, benefit the collective but do not generate any immediate financial returns to the actors. Such behaviors are often referred to as *pro-social*, because social rewards have been identified as an important incentive. One important aspect of social reward, which we believe is crucial in the context of a social broadcasting network, is reputation. Reputation involves an estimation of one's skills (Jones et al. 1997) and perceived reputation enhancement has been identified as an important factor in motivating sharing in the literature of information systems and management (Cheung and Lee 2012; Kankanhalli et al. 2005; Wasko and Faraj 2005). For example, it is argued that the importance of reputation is increasing in most organizations today and knowledge contributors can benefit from improved reputation (Kankanhalli et al. 2005). Focusing on a consumer review community that is much looser in its organization structure than traditional organizations, Cheung and Lee (2012) also identified reputation as an important factor significantly related to consumers' actions of spreading electronic word-of-mouth. In social broadcasting networks, one's skill in filtering large amounts of content and digging out valuable pieces is highly appreciated by peers because the primary value of social broadcasting networks relies on their role as a platform for information provision and consumption (Bakshy et al. 2011; Kwak et al. 2010). Users achieve "success" on such a platform by consistently providing quality information. Sharing novel content created by others enhances a user's reputation as a connected person and as a person capable of filtering content and identifying valuable information. Therefore, after receiving a piece of content containing novel information, one has the motivation to forward the content to his or her followers in the social broadcasting network. The strength of that motivation depends on the extent to which the receiver's reputation can be enhanced.

From a potential sharer's perspective, how far his or her reputation can be enhanced by sharing a message is determined by two factors: the number of subscribers who would receive the shared content and the extent to which those subscribers value that piece of content. The subscribers' valuation of the content depends partly on the intrinsic quality of the shared information: The higher the quality, the more the audience values the content—hence, the greater the latent benefit of sharing is to the sharer.[15] Moreover, valuations of the same content (quality) by different audiences should also differ because each has different preferences and different knowledge sets. For instance, the early tweet about the death of Osama bin Laden should indeed have had high informational value to most ordinary Twitter users. However, for anyone inside the White House Situation Room on May 1, 2011, that tweet simply repeated a story he or she already knew and thus was of little additional informational value. This distinction suggests that information consumers with different backgrounds could attach unequal values to the same piece of content. In particular, the novelty of the information should affect a particular consumer's valuation. Apparently, the degree to which a potential sharer will perceive a piece of content as novel to followers depends on how deeply the sharer actually regards the information as novel.[16] Earlier

[14] The cost could be interpreted as a capacity constraint or the opportunity cost of choosing not to share.

[15] Because of this quality effect, we cluster our observations based on each tweet in our analysis.

[16] Another possible factor is the extent to which the sharer believes his or her followers have already been exposed to the content. This degree may be captured by overlapping the sharer's followers and the content creator's followers. We take this into account in an extension of our empirical study in the subsection "Theoretical Motivation Revisited."

works in sociology suggested that the strength of the social tie between the sender (i.e., content creator) and the receiver (i.e., potential sharer) of a piece of information is closely related to how much the receiver would perceive the information as novel. For example, Granovetter (1973) theorized about the relationship between the novelty of information and the strength of the social tie through which the information is transmitted in the context of people finding jobs. Granovetter's results suggested that weak ties—those personal connections linking distant acquaintances—were more likely to provide nonredundant information because strong ties link closely related persons, such as family and friends, who often possess knowledge sets similar to the job seeker's. Following Granovetter's seminal work, subsequent research further demonstrated that, in both real organizations and virtual communities, weak ties are critical in connecting diverse groups and enabling a person to access heterogeneous and thus more novel information (see Constant et al. 1996; Granovetter 1983; Hansen 1999; Levin and Cross 2004; Weenig and Midden 1991). Adopting this view in the context of information sharing in a social broadcasting environment, we hypothesize that the strength of the social tie between the content creator and a potential sharer mediates the sharer's belief about the latent benefit the sharer may obtain by sharing the content with his or her followers. Specifically, *on average*, the weaker the tie, the higher a potential sharer believes followers would value the information and hence the higher the expected reputation enhancement if the potential sharer chooses to forward the content to his or her followers. The implication of this line of argument is the following hypothesized relationship between content-sharing probability and tie strength:

> **Hypothesis 1**: *In social broadcasting networks, the latent benefit of sharing content is negatively associated with the strength of the social tie between a potential sharer and the content creator. Thus, given a piece of content, a weak-tie subscriber is more likely to share than a strong-tie subscriber, everything else being equal.*

It is important to compare our hypothesis with SWT. A key implication of SWT in the context of a social broadcasting network is the expectation that followers of a potential sharer who has a weak tie with the creator of the content will usually attach a higher value to the content. For this very reason, we argue, the potential sharer will have a larger incentive to forward the content to his or her followers in anticipation of obtaining a higher reputation enhancement. Although our hypothesis is closely related to SWT and even resembles SWT in its form, it is neither a simple repetition nor a straightforward application of it. SWT states only that information obtained from one's weak-tie connections is expected to be more valuable; it says nothing about whether weak ties actually promote information dissemination. These two ideas are fundamentally very different concepts.[17] Hence, our hypothesis extends the original SWT findings within the theoretical framework of the social exchange by arguing that in social broadcasting networks, weak ties, in expectation of higher social exchange returns, are more likely to become the paths by which information is relayed.

User relationships in the Twitter environment are apparently not exactly the same as the real-world personal relationships on which Granovetter initially focused to study the strength of weak ties. Hence, to adapt our hypothesis in the Twitter world and test it with data, we need to operationalize empirically the strength of social ties in the Twitter network. We do this step based on the observed relationship types and assume that reciprocal relationships are *on average* stronger than nonreciprocal ones. This assumption leads to the following assumption, which is key to our subsequent empirical analysis:

> **Assumption**: *A unidirectional link between two Twitter users is expected to be weaker than a bidirectional one, in the sense of tie strength established by Granovetter (1973).*

For instance, in the example of Figure 1, ties like *D–R* are expected to be weaker than those like *C–R*.

Even though our measure of tie strength is very natural, it nonetheless needs to be supported by convincing theoretical arguments and empirical evidence. We provide the supporting argument of our assumption in Appendix A for interested readers. Meanwhile, we merely note here that the emphasis on reciprocity is consistent with a long tradition in the sociology literature. For example, Davis (1970) suggests that mutual choices indicate a strong tie while asymmetric pairs indicate weak ties.[18] Granovetter (1973) also pointed out that the strength of a tie is a combination of several factors, including mutual confiding and reciprocal services. Friedkin (1980) measured tie strength among faculty members in seven biological science departments of a single university based on whether a discussion about current research is reciprocated or not reciprocated.

[17] We thank an anonymous reviewer for making this point.

[18] Davis measured interpersonal relations on a three-point ordinal scale: mutual positives are the most positive, mutual negatives are least positive, and asymmetric pairs are intermediate. In sociometry, these correspond to mutual choices (*i* chooses *j* and *j* chooses *i*), mutual nonchoices (*i* does not choose *j*, and *j* does not choose *i*), and unreciprocated (*i* chooses *j* but *j* does not choose *i*, or *j* chooses *i* but *i* does not choose *j*).

Based on the assumption, our hypothesis, adapted in the Twitter world, becomes an empirically testable one:

> **Hypothesis 2**: *On expectation, a unidirectional follower is more likely to retweet than a bidirectional follower.*

For instance, in Figure 1, *ex ante* we expect *D* is more likely to retweet *R*'s tweet than *C* is. We develop our econometric model based on both of these theoretical discussions and the technological specifics of the Twitter environment. Before discussing the model, we describe our data.

Data

We deployed 20 computers to collect data by querying Twitter's application programming interface (API).[19]

Data Collection

Figure 2 shows the data collection workflow and is a useful illustration for helping readers to understand the details of our data collection process, described in the following paragraphs. From July 22, 2010, to December 2, 2010, at 0:05 each day, our "pick-tweet" program fetched Twitter's *toptweets* web page, which usually showed 17 to 18 popular tweets in the Twittersphere at the visiting time.[20] Sorting these tweets into chronological order, our program then checked, one by one, the number of followers a tweet's author had and inserted into our tweets database the first one it found whose author had less than 1,500 followers; the rest were discarded. If all the authors had more than 1,500 followers, the program wouldn't insert any tweet on that day. In other words, our program picked either 1 tweet or 0 tweets every day over this period of time.[21]

After a tweet entered our tweets database, another "fetch-retweeter" program began to track and fetch its retweeting data and would do so constantly during the subsequent five days.[22] At 10 minutes past each clock hour over the 5 days, the program queried Twitter API to get the user IDs of the retweeters (those in shaded circles in the Figure 1 example). The retweeter IDs were obtained in the order of the time at which the user retweeted.[23]

As retweeting data came in, another "fetch-graph" program worked on collecting relevant network graph information. Specifically, for each tweet, we were interested in its author (*R* in Figure 1), the author's followers (*A*, *B*, *C*, *D*, and *E* in Figure 1), and the tweet's retweeters (*A*, *D*, and *E* in Figure 1); we called this set of Twitter users our focal set. For each user in the focal set, our program collected the IDs of both the user's followings and followers and stored the data in our network graph database. For some users in the focal set, access to their following-IDs and follower-IDs was restricted because they explicitly disallowed third-party access to their data. We used a "protected" flag to indicate this privacy protection status, with the flag meaning no public data access. With the retweeting data and network graph data in hand, we produced a real-world analog of Figure 1 (see Appendix B). The figure shows the spread of the first tweet in our database.

We designed our data collection strategy around one important binding constraint: Twitter API allowed only 150 visits/queries per IP per hour,[24] and our computer and network resource was limited. One API visit would return only a limited amount of information, so to finish one "job" (e.g., getting the entire set of a user's following-IDs) could require a number of queries (e.g., the actual number of visits required

[19] See http://dev.twitter.com.

[20] *Top Tweets* is an official Twitter account, which is a "new algorithm that finds tweets that are catching the attention of other users." The algorithm is proprietary, so we cannot give a definition for a "popular tweet." Twitter's chief scientist, Abdur Chowdhury, explained, "The algorithm looks at all kinds of interactions with tweets, including retweets, favorites, and more to identify the tweets with the highest velocity beyond expectations."

[21] The "pick-tweet" program did not run properly on a few days during our data collection period because technical problems (e.g., server failure) occurred on either the Twitter side or our side. On those days, no tweets were added to our database.

[22] The decision to track retweeting activities for five days was made on the basis of our judgment about how long a retweeting process of one tweet could stay active. The log file written by the "fetch-retweeters" program showed that most retweeting activities of a tweet happened within just one or two days of when it was first posted. Tracking for five days thus seemed conservative enough to ensure that any truncated sample problem (a large number of retweets occurring after our tracking period) was unlikely.

[23] One important technical constraint was that Twitter API provided IDs for only the 800 most recent retweeters, so that if more than 800 users retweeted a tweet between two queries, our program was not able to get the complete set of retweeters. In addition, we found no publicly available way to verify the number of retweeters our program had missed. We took a conservative approach to deal with this situation: Unless we were sure we had fetched the complete set of retweeters for a tweet, we discarded that tweet from our database.

[24] This REST API rate limit was as of the second half of 2010 (https://dev.twitter.com/docs/rate-limiting).

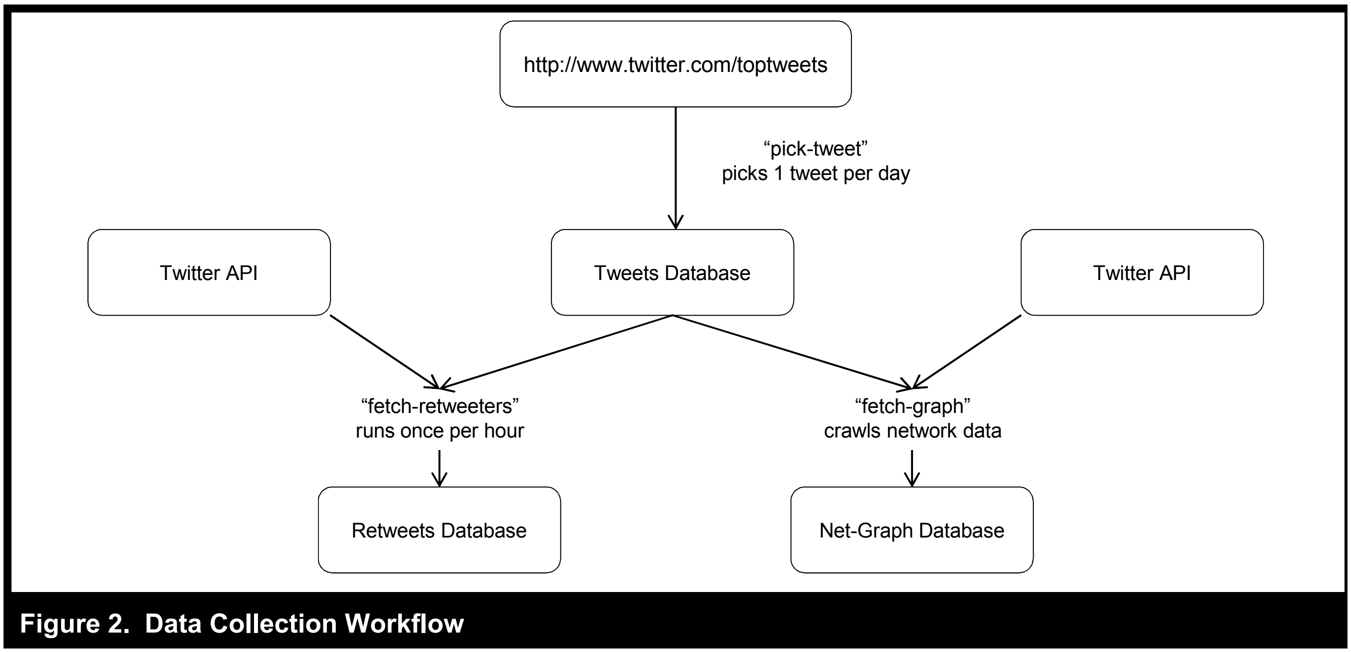


Figure 2. Data Collection Workflow

would depend on the number of followings the user had). As discussed in the previous paragraph, we had to collect all following-IDs and follower-IDs for all users in the focal set; moreover, we had to finish collecting the data as quickly as possible to avoid potential significant changes in their following–follower relationships. This 150-visits limit was the reason why we decided to select only one tweet per day, select only tweets whose authors had fewer than 1,500 followers, and track retweeting activity only once per hour, and why we decided *not* to collect network graph data of followers' followers (*G* in Figure 1).[25] Deciding otherwise would have prevented us from finishing the workload for one tweet before the next tweet was entered in our database.[26]

Data Description and Statistics

We provide a list of notations in Table 1.

Tweets, Authors, and the Number of Observations

By the end of the 140-day data collection course, we had successfully completed data collection for 65 tweets. We index the tweets in order of posting time by an integer, t, ranging from 1 to 65. The tweets were all authored by different users, so we also denote the author of tweet t author t, for simplicity of notation.[27]

[25] As a result, we do not have the second-order retweeters' network characteristics and we do not include the second-order (or the higher-order) retweeters in later econometric analyses. However, we believe this process is not a severe limitation of our study. Since a second-order retweeter must be a follower of one of the first-order retweeters and a retweet is a rebroadcast, we can think of the retweet as originating from the first-order retweeter and apply a similar analytical procedure. Of course, we cannot conclude that the higher-order user's behavior is the same as the immediate followers' behavior without any empirical evidence. Studying the similarities and differences could be a future research topic. In Appendix C, we provide a table detailing the number of retweets by immediate followers, second- or higher-order followers, and the other users for the tweets in our sample.

[26] An ideal situation is to collect a dataset by randomly sampling tweets from the entire Twittersphere without imposing any restrictions. However, most tweets do not generate any retweets. Using tweets published on Top Tweets (@toptweets) solves this problem, but it may raise concern over the generalizability of our findings. We argue that this situation is not as severe as it appears to be, since the variation in data we exploit is the systematic difference in unidirectional and bidirectional followers' reactions to *the same tweet*. We make this point clear in the empirical model section. In addition, we also collect a random sample of tweets, for which we provide statistics that can be compared with the Top Tweets sample. The results are shown in Appendix D.

[27] Among the 65 tweets, 3 are in Spanish, 1 is in Italian, 1 is in Portuguese, and the remaining are in English. None of the authors is a celebrity, partly because of our 1,500-follower constraint. The textual contents range from breaking news and comments on news to political jokes and witty quotes.

Table 1. Notations

Tweet level	t	index of tweets/authors
	n_t	the number of followers of author *t*, also the number of observations for tweet *t*
	v_t	the total number of retweeters of tweet *t*
Follower level	ti	index of author *t*'s followers, $i \in \{1, 2, \ldots, n_t\}$
	y_{ti}	binary outcome, = 1 if follower *ti* retweeted tweet *t*
	w_{ti}	binary variable, = 1 if follower *ti* is a unidirectional follower of *t* (weak tie)
	V_{ti}	the number of *ti*'s followings
	W_{ti}	the number of *ti*'s followers
	m_{ti}	the number of times *ti*'s followings retweeted tweet *t* (before *ti* did if y_{ti} = 1)

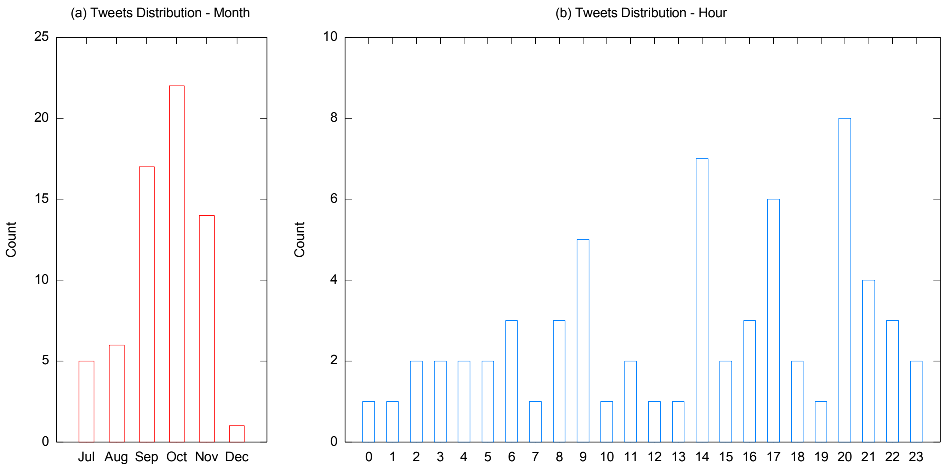


Figure 3. Distributions of Tweets by Month of Post and by Hour of Post

The two plots in Figure 3 show the distributions of the tweets by month of post (a) and by hour of post (b), respectively. The sample frequency of tweets by hour of post is roughly consistent with the distribution of total volume of tweets posted in each clock hour in the entire Twitter world. Subplot (a) of Figure 4 (the one on the left) shows two histograms of the number of tweets in the range defined by the number of followers the original author had (n_t) (unshaded bars) and the *total* number of retweeters the tweet gained (v_t) (shaded bars). Coincidently, the maximum total number of retweeters is also smaller than 1,500 (1,479). Note that for a tweet, v_t could be larger than n_t because retweeters' followers who were not immediate followers of the author could also have retweeted. The right subplot (b) of Figure 4 is a scatter-plot of the 65 tweets on the n_t–v_t plane. More or less surprisingly, our sample shows no positive correlation between the number of followers an author had and the total number of retweeters her tweet gained (a linear fitting line shows a weakly negative slope). However, this simple result is actually consistent with the study by Bakshy et al. (2011), which also finds that the number of an author's followers is in general a poor predictor of the size of the retweet cascade.

Because our objective is to model a follower's binary decision of whether to retweet, n_t, the number of followers that author *t* had is also the number of observations in cluster *t*. From this point onward, we exclude users for whom we could not collect following/follower IDs (flag "protected" = 1) and users with zero following/followers (assuming they were either new registrants or inactive members). As a result, the total number of observations ($N = \sum_{t=1}^{65} n_t$) in our sample declined from 29,681 to 24,403, a decrease of 17.78 percent. Table 2 gives the basic descriptive statistics of n_t before and after dropping the observations, and Figure 5 shows the number of pre-dropping versus post-dropping observations in more detail.

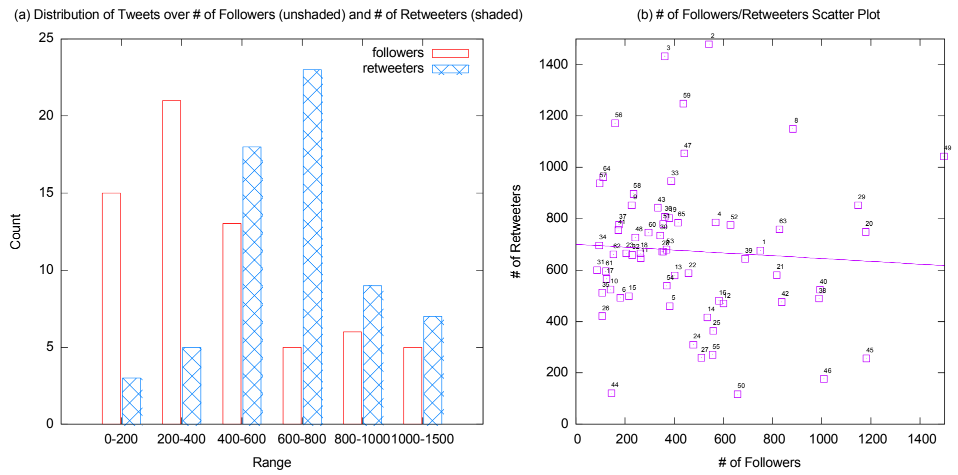


Figure 4. Distribution of Number of Author's Followers and Number of Retweeters

Table 2. Number of Observations per Tweet

n_t	*min*	*max*	*mean*	*median*
Total	87	1,497	467	370
Non-protected	54	1,189	375	324

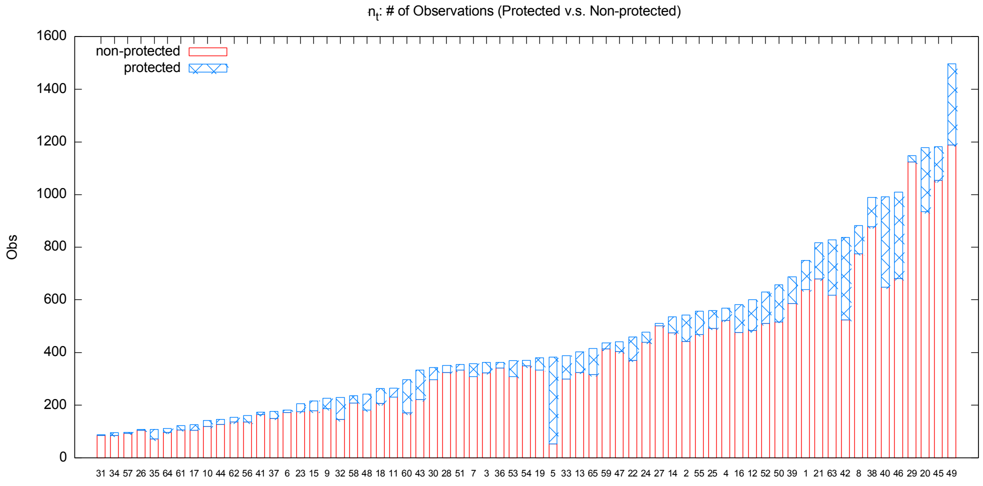


Figure 5. Number of Observations per Tweet

Variables

We now summarize the key variables used in the econometric model. For a tweet *t*, we use y_{ti}, $i \in \{1, 2, \ldots, n_t\}$ to index whether each of its observations (i.e., author *t*'s followers) retweeted tweet *t*. The definitions of the key variables can be found in Table 1. These variables are either directly observed or constructed from observed ones. We provide the descriptive statistics of these variables in Table 3 and the correlations between them in Table 4.

Table 3. Descriptive Statistics

		mean	*std.*	5%	15%	50%	85%	95%
y_{ti}	retweet	0.0427	0.2022	–	–	–	–	–
w_{ti}	unidirectional	0.7598	0.4272	–	–	–	–	–
V_{ti}	followings	1,574	9,046	25	69	347	1,714	3,297
W_{ti}	followers	3,304	73,124	5	22	190	1,117	4,970
m_{ti}	repetition	3.2845	7.5216	1	1	1	4	11

Table 4. Correlations

		y_{ti}	w_{ti}	V_{ti}	W_{ti}	m_{ti}
y_{ti}	retweet	1.0000				
w_{ti}	unidirectional	0.0072	1.0000			
V_{ti}	followings	-0.0225	-0.0921	1.0000		
W_{ti}	followers	-0.0065	-0.0338	0.4436	1.0000	
m_{ti}	repetition	0.0493	-0.1508	0.2002	0.1400	1.0000

Let $y_t = \sum_{i=1}^{n_t} y_{ti}$ be the number of retweeters among author t's followers (note that $y_t \neq v_t$), and $yr_t = y_t/n_t$ could then be naturally interpreted as the retweeting rate of t. Figure 6 shows the retweeting rate across the tweets with a 95 percent error bar. That the rate varies quite a lot is not surprising given the significant heterogeneity across the tweets (i.e., the intrinsic quality). Hence, we should consider tweet-specific effects when modeling retweeting behavior. Over the whole sample (i.e., tweets pooled together), the retweeting rate is 0.0427, and the 95 percent confidence interval is (0.0402, 0.0452).[28]

w_{ti} is the binary indicator of unidirectional relationship, which is also our main operationalization of a weak tie in econometric analysis. The simple correlation of y_{ti} and w_{ti} is positive. $w_t = \sum_{i=1}^{n_t} w_{ti}$ is the number of author t's followers who were not followed back by t. $wr_t = w_t/n_t$ is thus the fraction of t's unidirectional followers. We plot wr_t in Figure 7, which shows that for most of the tweets in our sample, wr_t is in the range (0.5, 0.9). Over the whole sample, the fraction is wr = 0.7598, and its 95 percent confidence interval is (0.7545, 09.7652).[29]

Some basic descriptive statistics of the number of followings (V_{ti}) and the number of followers (W_{ti}) can be found in Table 3. The *median* values of both V_{ti} and W_{ti} are much smaller than their respective *mean* values, so both distributions are positively skewed and have long right tails (i.e., the majority of the users had tens or hundreds of followings and followers, but a handful of them might have had up to hundreds of thousands of followings or even millions of followers). Similar statistics can be found in Kwak et al. (2010) and Wu et al. (2011), but the *median* numbers are much larger in our study than in their articles because we exclude observations with zero followings/followers. The Pearson's correlation of V and W is 0.4436, as shown in Table 4, and both V and W are negatively correlated with y_{ti}.

m_{ti} is the number of times someone among ti's followings (re)tweeted t (including author t's original tweet). m_{ti} also has a heavily positively skewed distribution: more than half of the observations received the tweet just once (i.e., none of their followings retweeted). Over the whole sample, the mean is equal to 3.28, and the standard deviation is equal to 7.53. We observe that m_{ti} is positively correlated with V_{ti} (the number of followings a user has) because m_{ti} is by definition

[28] Because we selected popular tweets, this retweeting rate does not generalize to the entire tweet space.

[29] We also compute the fraction of unidirectional links among all 110,583,366 relationships observed in our database (not only those between authors and their followers); the percentage is 75.2%, which is surprisingly close to wr. In other words, this finding says that, on average, roughly one out of four edges in the Twitter world is bidirectional. Kwak et al. (2010) crawled the entire Twitter network in July 2009 and computed this rate to be 77.9%; thus we see a higher fraction of bidirectional links one year after their research. This increment might be an interesting metric for researchers who study network formation.

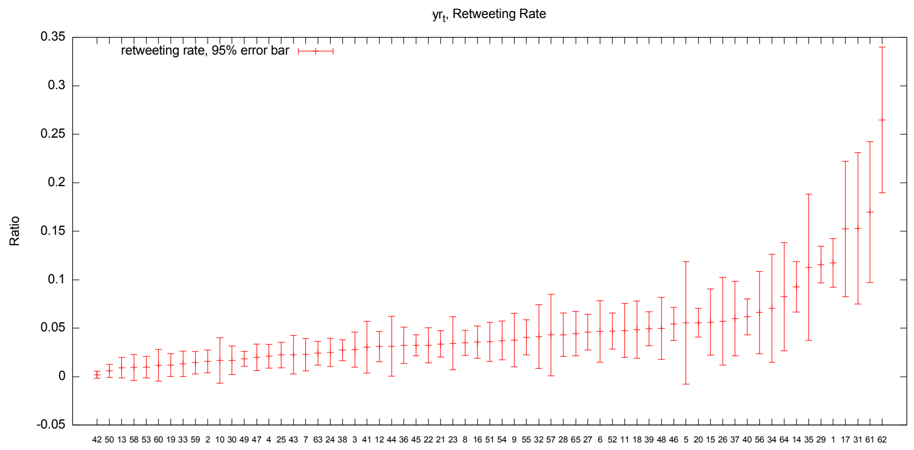


Figure 6. Retweeting Rate Across Tweets

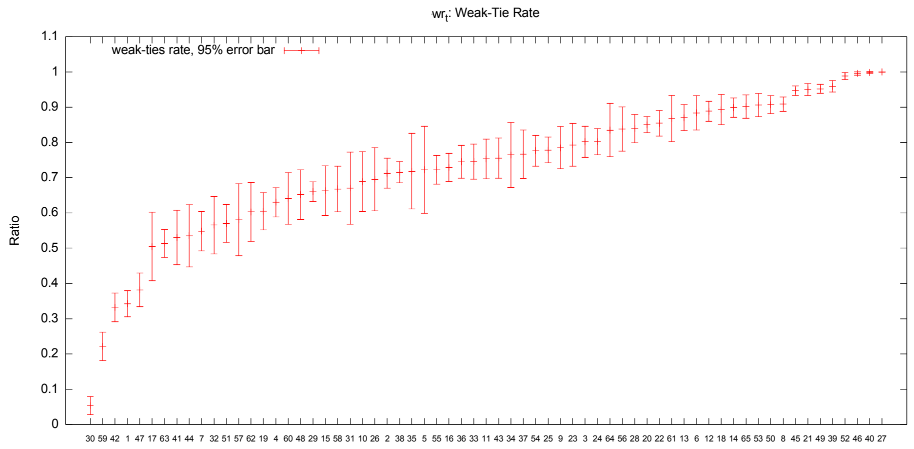


Figure 7. Weak-Tie Rate Across Tweets

the size of a subset of followings. m_{ti} is negatively correlated with w_{ti}, meaning bidirectional followers are likely to receive more retweets than unidirectional ones.

Empirical Model and Results

In this section we use our retweet dataset to perform empirical tests on our hypothesis. Instead of using standard reduced-form econometric methods for binary response (e.g., probit or logit), we take a more structural approach, modeling both the user behavior and special features of social broadcasting technology. We then use the MLE technique to estimate the empirical model and present the results.

Conditional MLE

We model a two-stage, consumption–retweeting process, in which consumption is the necessary first step for retweeting. We first describe the two stages and derive the likelihood function that would be used in our final conditional MLE analysis. We then show the results and discuss our findings.

Stage One: Consumption

The first stage models whether a follower of author *t*, say, *ti*, after receiving a tweet, actually consumes it. Figure 8 illustrates the technological aspect of this stage. The horizontal line stands for *ti*'s *home time line* or *Twitter feed*, which is a

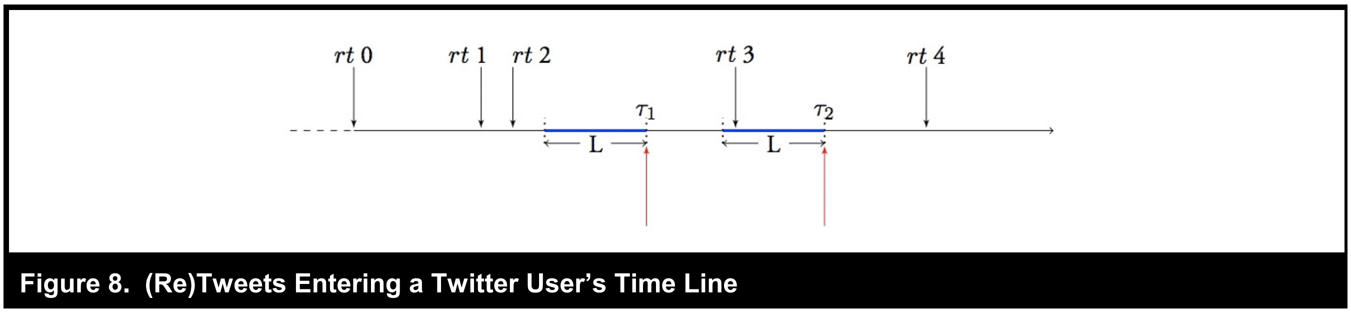


Figure 8. (Re)Tweets Entering a Twitter User's Time Line

stream of received tweets for *ti* to consume (read), including retweets, listed in chronological order. Note that not only the original tweet *t* but also *ti*'s followings' retweets of it, if any, appear in *ti*'s time line. The downward pointing arrows show the times at which a total of five (re)tweets of *t* enter the feed. Between these five (re)tweets, other tweets are also posted by *ti*'s followings.

In reality, few Twitter users can or will monitor their Twitter feed continuously. We assume every time they start reading their feeds, they consume only a limited number of tweets. In the example shown in Figure 8, the upward pointing arrows indicate the times, τ_1 and τ_2, when user *ti* launches her Twitter application. Because the tweets are listed in chronological order, tweets posted at times close to the τs are more likely to be consumed. For simplicity, we use a thick horizontal segment to indicate a "period of attention" of length *L*, inside which tweets posted are consumed. In doing so, we implicitly assume that users do not discriminate between tweets authored by different people. The only factor determining whether a tweet catches the user's attention is whether it enters the time line during a certain period preceding the time a user checks tweets.[30]

Therefore, the cognitive limit restricts a user *ti* from reading every single tweet she receives. In Figure 8, tweets that enter into the time line in the interval $(\tau_1, \tau_2 - L)$ are outside any of the periods of attention and would not be consumed by *ti*. When a tweet *t* gets retweeted by *ti*'s followings, it enters the time line multiple times, thus increasing the likelihood that *t* falls into one of the periods of attention (e.g., *rt*3 in the figure). If neither the original tweet *t* nor the retweets fall into some period of attention, then it is not consumed and hence would not be retweeted by *ti*.

Unfortunately, whether tweet *t* is actually consumed by *ti* is unobserved. Our task for this stage is to build a probabilistic model to capture the likelihood that *ti* consumes *t*, conditional on observed variables. Based on previous discussions about the technology, whether *ti* consumes tweet *t* is determined by three factors: (1) m_{ti}, the number of times *t* appears in *ti*'s time line; (2) the frequency with which *ti* checks her Twitter feed; and (3) *L*, which is determined by the number of tweets *ti* can read in each consumption and the number of tweets *ti* receives per unit of time, which we assume to be a linear function of V_{ti} (i.e., the more people a user follows, on average, the more tweets she receives over a fixed time span). Therefore, we propose the condition for *ti* to consume *t* be the following equation:

$$\frac{m_{ti}}{bV_{ti}} > a_{ti} \tag{1}$$

where *b* is a positive constant and $1/(bV_{ti})$ measures *L*.[31] The unobserved variable a_{ti} can be interpreted as an inverse measure of the frequency with which *ti* checks her Twitter feed, and is assumed to be independent of both V_{ti} and m_{ti}. The left side of equation (1) can be seen as the scaled frequency with which *t* appears in the time line, and the right side as a user-specific threshold. If a user does not check her feed very often, so that she gets a high draw of a_{ti}, then the scaled frequency needs to be high for the tweet to be consumed, and vice versa. To derive the likelihood function, we further assume that a_{ti} is log-normally distributed in the population:

$$\log a_{ti} | t \sim \log a_{ti} \sim N\left(a, \sigma_a^2\right) \tag{2}$$

So we can rewrite equation (1) as

[30] This random-reading modeling assumption is only a rough approximation of the real consumption stage. In reality, great variation exists in how people use Twitter and read their Twitter feed. However, because most people receive a large amount of tweets, of which they are able to consume only a portion, we believe that without detailed data on individual Twitter usage, random-reading is an appropriate modeling approximation for us to use.

[31] Or more generally, we can assume $L - \frac{z_{ti}}{bV_{ti}}$, where z_{ti} is the number of tweets *ti* can read in each consumption and bV_{ti}, $b > 0$, is the number of tweets received by *ti* per unit of time. We can still get (1) by dividing both sides by z_{ti} and absorbing the unobserved z_{ti} into a_{ti}.

$$\begin{aligned} -\log b + \log m_{ti} - \log V_{ti} \quad & > \log a_{ti} \\ -\tfrac{a=\log b}{\sigma^2} + \tfrac{1}{\sigma^2}\log m_{ti} - \tfrac{1}{\sigma^2}\log V_{ti} \quad & > \tfrac{\log a_{ti}-a}{\sigma^2} \end{aligned}$$

where the term on the right side is a standard normal distribution. Thus, the *ex ante* probability that *ti* consumes tweet *t*, conditional on receipt, is

$$\begin{aligned} p_1 &= p\left(-\tfrac{a+\log b}{\sigma_a} + \tfrac{1}{\sigma_a}\log m_{ti} - \tfrac{1}{\sigma_a}\log V_{ti} > \tfrac{\log a_{ti}-a}{\sigma_a}\right) \\ &= \Phi\left(-\tfrac{a+\log b}{\sigma_a} + \tfrac{1}{\sigma_a}\log m_{ti} - \tfrac{1}{\sigma_a}\log V_{ti}\right) \end{aligned} \tag{3}$$

where Φ is the cumulative distribution function (CDF) of the standard normal distribution. The outcome of this stage is unobserved, so we cannot estimate the parameters in which we are interested just on the basis of equation (3).

Stage Two: Retweeting

Recall that a follower *ti* retweets only if *ti* consumes the tweet himself. If a user's first stage outcome is a failure (he does not consume *t*), then his final outcome would automatically be *not retweeting*, $y_{ti} = 0$. In other words, $y_{ti} = 1$ implies success at both stages. Unlike the first stage, where success is determined by the broadcasting technology and chance, the second stage outcome depends on the decision made by the user.

At the second stage, the users who have consumed the tweets each decide whether to retweet. The decision is made on the basis of a subjective cost–benefit analysis. As discussed earlier, the latent benefit of retweeting depends on both the number of followers to whom the content is retweeted, W_{ti}, and the mean valuation the followers attach to the tweet, which we denote α_{ti}. Thus, we write the latent benefit $\alpha_{ti}W_{ti}$. We expect *ti*'s followers' mean valuation, α_{ti}, to be moderated by the strength of the social tie connecting author *t* and potential retweeter *ti*. Finally, for the retweeting act to happen, the latent benefit should exceed the user-specific reservation utility or cost, denoted c_{ti}. Therefore, after using logarithmic transformation, the necessary and sufficient condition of retweeting upon consumption can be written (with a slight abuse of the notation *a* and *c*):

$$\alpha_t + \delta w_{ti} + \beta \log W_{ti} > c_{ti} \tag{4}$$

where c_{ti}, like a_{ti}, is unobserved, and α, subindexed by *t*, is allowed to differ across the tweets, capturing tweet-specific effect.[32]

Technically, we further assume c_{ti} is distributed normally among the population. We also allow the unobservables at the two stages to be correlated:

$$c_{ti}|t \sim c_{ti} \sim N(c, \sigma_c^2),\ \mathrm{Cor}(c_{ti}, a_{ti}) = \rho \tag{5}$$

We can rewrite equation (4) as

$$-\tfrac{c}{\sigma_c} + \tfrac{\alpha_t}{\sigma_c} + \tfrac{\delta}{\sigma_c} w_{ti} + \tfrac{\beta}{\sigma_c}\log W_{ti} > \tfrac{c_{ti}-c}{\sigma_c}$$

where the right side is a standard normal distribution. Therefore, the conditional probability of retweeting can be written as follows:

$$p_2 = \mathrm{p}\begin{pmatrix} -\tfrac{c}{\sigma_c} + \tfrac{\alpha_t}{\sigma_c} + \tfrac{\delta}{\sigma_c} w_{ti} + \tfrac{\beta}{\sigma_c}\log W_{ti} > \tfrac{c_{ti}-c}{\sigma_c} | \\ -\tfrac{\alpha+\log b}{\sigma_a} + \tfrac{1}{\sigma_a}\log m_{ti} - \tfrac{1}{\sigma_a}\log V_{ti} > \tfrac{\log a_{ti}-a}{\sigma_a} \end{pmatrix} \tag{6}$$

Two-Stage Model for MLE

At this point, we put the two stages together. Equations (2), (3), (5), and (6) represent all of the necessary elements for conducting the MLE analysis. The likelihood of observing outcome $y_{ti} = 1$ for tweet *t* and follower *ti* is the product of p_1 and p_2, and the likelihood of observing $y_{ti} = 0$ is $1 - \mathrm{p}(y_{ti} = 1)$. In terms of econometrics, not all of the structural parameters are identified. For example, we can identify δ/σ_c, but not δ and σ_c separately. Fortunately, for our research purpose, we care most about the signs of the parameters rather than their absolute value. In the example, δ/σ_c has the same sign as δ; thus, identifying the ratio is good enough for understanding *w*'s partial effect. Therefore, for simplicity of notation, we rearrange the terms, rescale the parameters following the standard practices in probit and logit models, and obtain our benchmark specification:

$$\begin{aligned} &\mathrm{p}(y_{ti} = 1) = p_1 p_2 \\ &p_1 = \mathrm{p}(e + b_1 \log m_{ti} + b_2 \log V_{ti} > a_{ti}) \\ &p_2 = \mathrm{p}(\alpha_t + \delta w_{ti} + \beta \log W_{ti} > c_{ti} | e + b_1 \log m_{ti} + b_2 \log V_{ti} > a_{ti}) \\ &a_{ti}, c_{ti} \sim N(0,1) \\ &\mathrm{Cor}(a_{ti}, c_{ti}) = \rho \\ &\theta = \{e, b_1, b_2, \alpha_1, \alpha_2, \ldots, \alpha_T, \delta, \beta, \rho\} \end{aligned} \tag{7}$$

where θ is a vector of parameters to estimate. α_t (with *t* ranging from 1 to *T*) absorbs the constant term and captures the tweet-specific effects. δ is the coefficient of the weak-tie indicator, which is our primary interest. b_1, b_2, and β determine the partial effects of the other social network characteristic variables.

[32] α_t also includes the author-specific effect, since in our sample the tweets are all by different authors.

Results

With equation (7) in hand, we estimate the parameters using the conditional MLE method. We report the results in Table 5. We estimate a total of five different specifications, the first four of which are described in detail in the following paragraphs. The last one is discussed in the next subsection. In all specifications, we use dummy variables to capture tweet-specific effects,[33] α_ts, and we do not report these fixed effects because they are less interesting in our analysis.[34] All standard errors are computed to be robust to tweet clustering.

Model 1 corresponds to equation (7), with an additional restriction that $a_{ti} \perp c_{ti}$, which implies $\rho = 0$. Model 2 strictly follows the benchmark equation (7), allowing correlation between a_{ti} and c_{ti}. Models 3 and 4 slightly modify model 2: model 3 includes the interaction term of w_{ti} and W_{ti} in the retweeting equation; model 4 includes w_{ti} in the consumption equation.

We observe that the fitted likelihood increases from model 1, to model 2, and to models 3 and 4, as we gradually relax the model restriction by adding richer structures and more variables. Across the four columns, we find consistent support for a positive m_{ti} coefficient (repetition of retweets) and a negative V_{ti} coefficient (the number of followings). All estimates are significant with 99.9 percent confidence level. Therefore, the results are consistent with the model prediction described in this subsection and, in particular, with equation (3).

The unidirectional-relationship/weak-tie indicator is found to have a significantly positive effect on the (conditional) retweeting probability. In the benchmark model (model 2), its coefficient is positive at the 0.1 percent significance level. The w_{ti} coefficient becomes less significant, but is still positive at the 5 percent significance level, when we allow an interaction effect of tie-strength and the number of followers (model 3) or when we put the weak-tie indicator into both the consumption and retweeting equations (model 4). These results show that, in the retweeting equation, the positive sign of the weak-tie coefficient is robust; thus, they support our hypothesis: Weak ties are more likely than strong ties to relay information to their social network neighbors.

In model 3, where we include the weak-tie dummy w_{ti} in both the consumption and retweeting equations, we find that, although its effect on retweeting probability is positive and significant, its effect on consumption probability is negative but insignificant. This result shows that messages generated from stronger ties *might* be more likely to be read than those from weaker ties. However, the difference in likelihood is not statistically significant. It supports our assumption that users generally do not discriminate between tweets received from strong ties and tweets received from weak ties. We believe the separation of the different effects that weak ties have on the two probabilities, as model 3 reveals, shows the merit of our two-stage econometric model. It indeed uncovers more structure in the retweeting process than a reduced-form probit regression.

In all models, the number of followers has a significantly positive coefficient. This revelation by our econometric models is a new one because, as shown in Table 4, the simple correlation between y_{ti} and W_{ti} is negative. This result thus supports our argument in the theory section that the number of subscribers is positively associated with the latent benefit of retweeting.

Theoretical Motivation Revisited

From model 1 to model 4, we consistently find that, conditional on the consumption of a piece of information, weak-tie users are more likely to share information with their social network neighbors. In the theory section, we argued the reason is that a weak-tie follower's followers would *on average* value the information more than a strong-tie follower's followers; thus, the latent benefit from the social exchange of content sharing is greater for a weak-tie follower than for a strong-tie follower, everything else being equal.

In a social broadcasting environment, two possible explanations remain for the higher mean valuation of the shared content from a weak-tie follower's followers.

1. *New audience effect*: Because of the social broadcasting technology (in which whatever is posted or shared is broadcast to all followers), the possibility exists that the information has already been circulated to more of a strong-tie follower's followers than to a weak-tie follower's followers.[35] Holding the total number of a potential sharer's followers constant, the expected number of

[33] Technically, we can directly use dummy variables to control for fixed effects without appealing to more sophisticated econometric specifications because we have a large number of observations for every tweet. See Figure 5.

[34] We do not control for follower fixed effects because, for each tweet, all followers/observations are by definition distinct, and when we pool tweets together, among all the 24,403 observations, 24,002 are unique.

[35] One important observation is that a strong-tie follower's followers are more likely to be simultaneously following the author than a weak-tie follower's followers. Readers can refer to Appendix A for an empirical test.

Table 5. Results of Maximum Likelihood Estimation

Probability of Retweeting		**Model 1** Coeff. (-value)	**Model 2** Coeff. (-value)	**Model 3** Coeff. (-value)	**Model 4** Coeff. (-value)	**Model 5** Coeff. (-value)
		$p_{1:}$ Probability of Consumption upon Receipt				
log m_{ti}	Repetitions	0.494*** (4.81)	0.340*** (4.37)	0.340*** (4.37)	0.338*** (4.47)	0.434*** (5.86)
log V_{ti}	Followings	-0.639*** (-10.38)	-0.472*** (-5.14)	-0.473*** (-5.05)	-0.475*** (-5.02)	-0.566*** (-7.23)
w_{ti}	Weak tie				-0.076 (-0.46)	
		p_1: Probability of Retweeting upon Consumption				
w_{ti}	Weak tie	0.284*** (5.57)	0.220*** (5.52)	0.237* (2.14)	0.249** (3.22)	0.175*** (4.11)
log W_{ti}	Followers	0.115*** (6.32)	0.101*** (7.46)	0.103*** (4.95)	0.102*** (7.21)	0.131*** (6.83)
w_{ti} log W_{ti}	Weak tie × Followers			-0.003 (-0.17)		
$OI_{ti}^{W_1}$	Overlap Index of Followers I					-2.131* (-2.44)
$OI_{ti}^{W_2}$	Overlap Index of Followers II					-0.429 (-1.33)
ρ	Correlation (-value)	–	-0.836*** (0.000)	-0.835*** (0.000)	-0.834*** (0.000)	-0.606* (0.034)
Observations		24,403	24,403	24,403	24,403	24,403
Pseudo Log-Likelihood		-3,921.876	-3,913.125	-3,913.112	-3,913.010	-3,892.148

* = 0.1% significance level; ** = 1% significance level; *** = 5% significance level

followers who are new to the information is larger for a weak-tie follower. Therefore, a weak-tie follower can reach a larger new audience, and hence the sharing gives a greater social exchange benefit.

2. *Informational value effect*: In the absence of the *new audience effect* (the content is new to every follower), the content to be shared can still be *intrinsically* more valuable to a weak-tie follower's followers than to a stronger-tie follower's followers. The reason is that for the weak-tie follower's followers, the shared content comes from a relatively distant community, so it is more likely to complement their existing knowledge sets and hence be of higher informational value. Therefore, a weak-tie follower is more willing to share it because the sharing is expected to yield a higher social exchange benefit.

3. A third possibility is that both of these two effects exist.

We test the three possibilities in model 5 by adding two empirically constructed followers-overlap measures into the second-stage retweeting equation. Mathematically, we define two versions of an overlap index of followers:

$$OI_{ti}^{W_1} = \frac{\overline{W}_{ti}}{\sqrt{W_t\sqrt{W_{ti}}}}, OI_{ti}^{w_2} = \frac{\overline{W}_{ti}}{\min\{W_t, W_{ti}\}}$$

where $\overline{W}_{ti}$, W_t, and W_{ti} are the number of mutual followers author t and user ti shared, the number of followers author t had, and the number of followers ti had, respectively. $OI_{ti}^{W_1}$ and $OI_{ti}^{W_2}$ basically measure how "similar" user ti's followers and author r's followers are: The larger the index is, the more similar the two sets of followers are. The indexes are also used in Appendix A, where we test whether unidirectional relationships are weaker than bidirectional ones. Readers can refer to Appendix A for further discussion on the indexes.

We include $OI_{ti}^{W_1}$ and $OI_{ti}^{W_2}$ to capture the *new audience effect*, the first explanation. If it is indeed a driver of the result, we expect $OI_{ti}^{W_1}$ and $OI_{ti}^{W_2}$ collectively to have a negative effect on retweeting probability: If a user has a large number of followers who also follow the author, then he or she should be less willing to share the information. Moreover, if the *new audience effect* is the sole driver, then the weak-tie indicator w_{ti} should have no effect on retweeting probability once we include the two indexes.[36] If we find the two indexes have negative coefficients *and* the weak-tie indicator still has a positive coefficient, then we should conclude that both the *informational value effect* and the *new audience effect* exist.

The result of model 5 shows that the coefficients of the two indexes are indeed negative. Although the second version of the overlap index, $OI_{ti}^{W_2}$, separately is insignificant, collectively they are significant with a 99.9 percent confidence level. The magnitude of the coefficient of decreases from model 2, but it is still positive at 0.1 percent significance level. These two findings together support the third possibility: Both the *informational value effect* and the *new audience effect* exist.

Managerial Implications

Measuring a user's social influence in an online community is of great interest to managers who want to leverage the power of social media. On Twitter, a user is often regarded as being influential when many people retweet his or her tweets. Indeed, the depth of penetration and breadth of reach of one's words in an online community are important aspects of social influence. Our model measures the role that social network characteristics play in the information diffusion process on Twitter. Combined with the probability of consumption, we can compute the expected total number of consumers of a user's tweet based on his or her social network characteristics, which may serve as a starting point for measuring his or her social influence.

One important implication of our study is that having more followers does not directly translate into greater social influence.[37] In particular, the strength of social ties between a user and her followers should have an important moderating role, because it can greatly affect the followers' willingness to forward her messages. To see this more intuitively, we plot the fitted retweeting probabilities in Figure 9 for $w = 0$ (solid curve) and $w = 1$ (dashed curve), using estimates from model 2 and fixing α_t at the median value in our sample. The difference between the conditional probabilities of retweeting for a unidirectional follower and for a bidirectional follower is significant. For example, when W, the number of followers, equals 190, the median number in our sample, the conditional likelihoods of retweeting are 6.0 percent for a bidirectional follower and 9.1 percent for a unidirectional follower. The 3.1 percentage point difference means the latter is more than 50 percent higher in percentage.

However, the readers should be cautious in generalizing our results and the proposed method of measuring influence to non-broadcasting social networks. Our empirical operationalization depends on the existence of two types of interpersonal links on Twitter, so it is not directly applicable to undirected social networks such as Facebook. More importantly, the method proposed here results from our information diffusion model. Therefore, this method may not be the best for measuring social influence on social networks whose primary function is not spreading information (e.g., the professional social network LinkedIn).

Conclusion

An important question in the field of Information Systems is how information or knowledge is disseminated in an online community (with or without an organizational form). Large-scale empirical studies to address this question have traditionally been challenging because of the difficulty of obtaining detailed micro-level data. To the best of our knowledge, this paper is the first such study in the Information Systems field, where publicly available data from Twitter is used to explore people's voluntary information relay processes.

Using a carefully designed data collection process and a series of econometric analyses, we find that content is more likely to be relayed through weak ties on Twitter. This result is complementary to Granovetter's (1973) finding, which advocates for the important role of weak ties in carrying novel information. The implications of our finding are far-reaching. On the one hand, our theory, which is based on two highly influential sociological theories—the social exchange theory and the strength of weak tie theory—and is supported by the latest data from one of today's largest online social networks, reveals the important role that weak ties play in facilitating information dissemination in the social network through people's voluntarily information relay behavior. On the other

[36] Assume the two indices have perfectly captured the *new audience effect*.

[37] This view was recently shared by Twitter cofounder Evan Williams who hinted that follower counts may soon become the *second* most important number to users and the number of retweets is more interesting. For a full report, see http://www.buzzfeed.com/jwherrman/twitter-cofounder-suggests-a-replacement-for-the-f.

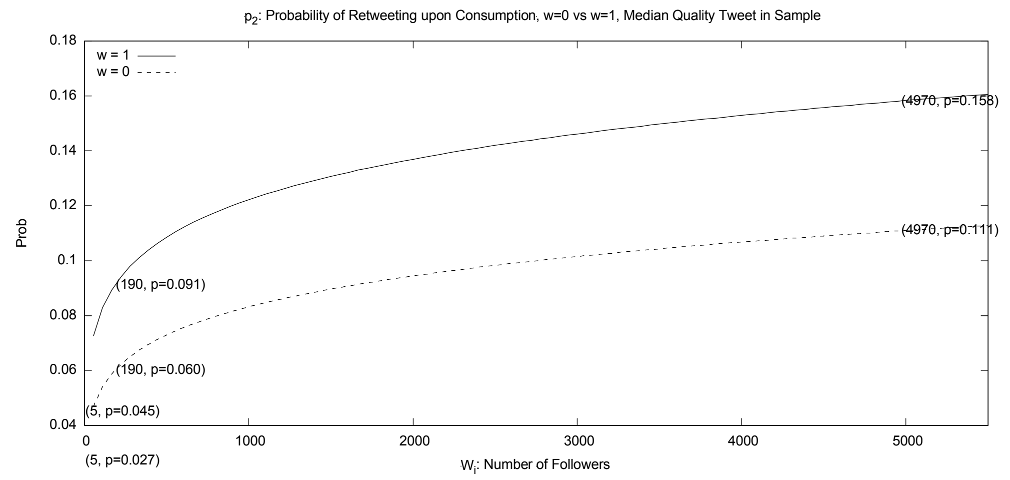


Figure 9. The Probability of Retweeting a Tweet Upon Consumption

hand, the interesting connection between tie strength and retweeting behavior indicates the importance of incorporating tie strength when measuring personal influence on Twitter, which is a question of fundamental importance to both researchers and practitioners.

As one of the first in the Information Systems field to bring together the huge amount of public data on Twitter with sociological theories to study information diffusion in social broadcasting networks, the paper is not without its limitations. First, the tweets in our data set were not randomly sampled. By using this data set to study the effect of tie strength in information sharing, we implicitly assumed that tweet "quality" changes everyone's retweeting probability only *uniformly*. Relaxing this assumption requires additional work (including obtaining a new data set) to test whether our results hold when the quality of tweets is moderate or low. Second, we measured tie strength using a binary variable based on whether a link is unidirectional or bidirectional. Measuring tie strength based on the amount of conversation between two Twitter users would be an alternative approach. Third, we used only an author's immediate followers and omitted higher-order potential followers in empirical analyses. As we discussed in the "Data" section, this was due to the difficulty of collecting network graph data for *all* higher-order potential retweeters. In future research, one could try to overcome the difficulty by, possibly, sampling these users. It is interesting to investigate the similarities and differences in sharing behavior between these higher-order retweeters and the immediate followers of tweets' authors. Fourth, we observed only one snapshot of the social network and thus modeled it as fixed and exogenous. Future research could examine the interplay of user behavior and the dynamics of underlying network structure. Another possibility for extending the current study is to include more user-specific variables (e.g., demographic information) and tweet-specific variables (e.g., constructed from natural language processing) into the econometric model. Of course, these extensions pose new challenges in terms of data collection and data processing. Nevertheless, they are certainly interesting directions to pursue in the future.

Acknowledgments

We thank the senior editor, Ravi Bapna, for his guidance and support throughout the review process. We thank the associate editor and the four anonymous reviewers for their constructive comments and concrete suggestions. We also thank Rion Snow from Twitter, as well as the other attendees to our panel at SXSW Interactive 2011, for useful comments and suggestions. The authors contributed equally to this paper.

References

Bakshy, E., Hofman, J., Mason, W., and Watts, D. 2011. "Everyone's an Influencer: Quantifying Influence on Twitter," in *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, Hong Kong, China, February 9-12, pp. 65-75.

Bapna, R., Gupta, A., Rice, S., and Sundararajan, A. 2012. "Trust, Reciprocity and the Strength of Social Ties: An Online Social Network Based Field Experiment," working paper, Carlson School of Management, University of Minnesota.

Blau, P. 1964. *Exchange and Power in Social Life*, Livingston, NJ: Transaction Publishers.

Bock, G., Zmud, R., Kim, Y., and Lee, J. 2005. "Behavioral Intention Formation in Knowledge Sharing: Examining the Roles of Extrinsic Motivators, Social-Psychological Forces, and Organizational Climate," *MIS Quarterly* (29:1), pp. 87-111.

Cheung, C. M. K., and Lee, M. K. O. 2012. "What Drives Consumers to Spread Electronic Word of Mouth in Online Consumer–Opinion Platforms," *Decision Support Systems* (53:1), pp. 218-225.

Chiu, C., Hsu, M., and Wang, E. 2006. "Understanding Knowledge Sharing in Virtual Communities: An Integration of Social Capital and Social Cognitive Theories," *Decision Support Systems* (42:3), pp. 1872-1888.

Constant, D., Sproull, L., and Kiesler, S. 1996. "The Kindness of Strangers: The Usefulness of Electronic Weak Ties for Technical Advice," *Organization Science* (7:2), pp. 119-135.

Davis, J. 1970. "Clustering and Hierarchy in Interpersonal Relations: Testing Two Graph Theoretical Models on 742 Sociomatrices," *American Sociological Review* (35:), pp. 843-851.

Friedkin, N. 1980. "A Test of Structural Features of Granovetter's Strength of Weak Ties Theory," *Social Networks* (2), pp. 411-422.

Friedkin, N. 1982. "Information Flow Through Strong and Weak Ties in Intraorganizational Social Networks," *Social Networks* (3), pp. 273-285.

Granovetter, M. 1973. "The Strength of Weak Ties," *The American Journal of Sociology* (78:6), pp. 1360-1380.

Granovetter, M. 1983. "The Strength of Weak Ties: A Network Theory Revisited," *Sociological Theory* (1), pp. 201-233.

Hansen, M. 1999. "The Search–Transfer Problem: The Role of Weak Ties in Sharing Knowledge across Organization Subunits," *Administrative Science Quarterly* (44), pp. 82-111.

Homans, G. 1958. "Social Behavior as Exchange," *The American Journal of Sociology* (63), pp. 597-606.

Jones, C., Hesterly, W. S., and Borgatti, S. P. 1997. "A General Theory of Network Governance: Exchange Conditions and Social Mechanisms," *Academy of Management Review* (22:4), pp. 911-945.

Kankanhalli, A., Tan, B. C. Y., and Wei, K. K. 2005. "Contributing Knowledge to Electronic Knowledge Repositories: An Empirical Investigation," *MIS Quarterly* (29:1), pp. 113-143.

Kwak, H., Lee, C., Park, H., and Moon, S. 2010. "What Is Twitter, a Social Network or a News Media," in *Proceedings of the 19th International Conference Companion on World Wide Web*, M. Rappa, P. Jones, J. Freire, and S. Chakrabarthi (eds.), Raleigh, NC, April 26-30, pp. 591-600.

Levin, D., and Cross, R. 2004. "The Strength of Weak Ties You Can Trust: The Mediating Role of Trust in Effective Knowledge Transfer," *Management Science* (50:11), pp. 1477-1490.

Lotan, G. 2011. "Breaking Bin Laden: Visualizing the Power of a Single Tweet," Socialflow, May 6 (http://blog.socialflow.com/post/5246404319/breaking-bin-laden-visualizing-the-power-of-a-single).

Olivera, F., Goodman, P., and Tan, S. 2008. "Contribution Behaviors in Distributed Environments," *MIS Quarterly* (32:1), pp. 23-42.

Wasko, M., and Faraj, S. 2005. "Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice," *MIS Quarterly* (29:1), pp. 35-57.

Weenig, M. W. H., and Midden, C. J. H. 1991. "Communication Network Influences on Information Diffusion and Persuasion," *Journal of Personality and Social Psychology* (61), pp. 734-742.

Wu, S., Hofman, J., Mason, W., and Watts, D. 2011. "Who Says What to Whom on Twitter," in *Proceedings of the 20th International Conference Companion on World Wide Web*, S. Srinivasan, K. Ramamritham, A. Kumar, M. P. Ravindra, E. Bertino, and R. Kumar (eds.), Hyderabad, India, March 28-April 1, pp. 704-714.

About the Authors

Zhan Shi is an assistant professor at the W. P. Carey School of Business, Arizona State University. Before joining ASU, he received his B.A. in Economics and B.S. in Mathematics from Peking University, Beijing, China, and his Ph.D. from the University of Texas at Austin. Since late 2010, he has been working at the Center for Research in Electronic Commerce, McCombs School of Business. His research focuses on analyzing user behavior in online social networks and understanding the impact of new social and mobile technologies on businesses and the economy.

Huaxia Rui is an assistant professor at the Simon School of Business, University of Rochester. Huaxia received his Ph.D in Information Systems in 2012 from the McCombs School of Business, University of Texas at Austin. His research interests include the study of social broadcasting network, display advertising, and securitization.

Andrew B. Whinston is Hugh Cullen Chair Professor in the Information, Risk, and Operation Management Department at the McCombs School of Business at the University of Texas at Austin. He is also the director at the Center for Research in Electronic Commerce and editor-in-chief of Decision Support Systems. His recent papers have appeared in *Information Systems Research*, *Management Science*, *Marketing Science*, *Journal of Marketing*, and *Journal of Economic Theory*. He has published over 300 papers in the major economic and management journals and has authored 27 books. In 2005, he received the Leo Award from the Association for Information Systems for his long-term research contribution to the information system field. In 2009, he was named the Distinguished Fellow by the INFORMS Information Systems Society in recognition of his outstanding intellectual contributions to the information systems discipline.

MIS Quarterly

RESEARCH ARTICLE

CONTENT SHARING IN A SOCIAL BROADCASTING ENVIRONMENT: EVIDENCE FROM TWITTER

Zhan Shi
Department of Information Systems, W. P. Carey School of Business, Arizona State University,
Tempe, AZ 85287 U.S.A. {zhan.m.shi@asu.edu}

Huaxia Rui
Simon Graduate School of Business, University of Rochester, Rochester, NY 14627 U.S.A. {huaxia.rui@simon.rochester.edu}

Andrew B. Whinston
McCombs School of Business, The University of Texas at Austin, Austin, TX 78712 U.S.A. {abw@uts.cc.utexas.edu}

Appendix A

Operationalization of Weak Ties

In this appendix, we discuss our operationalization of weak ties used in our empirical analyses. We define tie strength based on the following–follower relationships observed in the Twitter network, and, specifically, we claim that unidirectional relationships are *on average* weaker than bidirectional ones. We want to stress a few points regarding this assumption. First, we are not claiming that a bidirectional relationship in the Twitter world is a strong tie in the *absolute* sense. Twitter users, even if they are mutually connected online, often barely *know* each other in the real world, so to a certain extent, the claim that almost all ties on Twitter are weak is a fair one to make. The hypothesis only emphasizes the ordinal strength of the two types of ties, and the comparison is carried out in the sense of *probabilistic expectation*. The reason why reciprocity makes a difference is that frequent learning or regular interaction is more likely to happen when a reciprocal relationship exists. By reading each other's posts, a pair of users can more easily develop mutual understanding about each other's topics of interest and expertise, and sometimes even about detailed aspects of each one's personal life. Over time, even though the pair are unknown to each other in the real world, they might become very familiar with each other's activities and habits in the online community. Of course, reciprocal following does not guarantee such relationship development (which is why we emphasize the probabilistic nature of the hypothesis). However, without it, the relationship development is much less likely. Moreover, our operationalization is consistent with the previous sociological literature. Granovetter (1973) pointed out the importance of reciprocity by defining that "the strength of a tie is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie" (p. 1361). In Friedkin (1982), asymmetrical contact between college professors was classified as a weak tie, and a reciprocal connection was classified as a strong tie. Marlow et al. (2009) also applied similar definitions in analyzing friendships on Facebook.

We perform an empirical test on the hypothesis, using the network graph data we collected. Note that we know not only the number of followings (followers) a user has, but also who the followings (followers) are (i.e., we observe the IDs of the user's immediate social neighbors in our database). This information should give us more knowledge about, and in the meantime the ability to build important metrics of, a user's network characteristics. In particular, knowing the IDs of two users' social neighbors, we can compare how "similar" their social neighborhoods are. In deriving his theory, Granovetter, in his 1973 paper, claimed that the stronger the social tie between two persons, the larger the overlap of their friendship circles. Applying this statement in the Twittersphere, under our assumption, we would expect that two users who mutually follow each other, on average, have a larger overlap in their followings (followers) than two who don't. Our test is based on this prediction. Operationally, we do so by empirically verifying whether $w_{ti} = 0$ positively correlates with a higher similarity between user and author 's followings (followers). We measure similarity by computing two overlap indexes of followings (followers) of author *t* and user *ti*:

$$OI_{ti}^{V_1} = \frac{\overline{V}_{ti}}{\sqrt{V_t}\sqrt{V_{ti}}}, OI_{ti}^{V_2} = \frac{\overline{V}_{ti}}{\min\{V_t, V_{ti}\}} \tag{8}$$

where $\overline{V}_{ti}$, V_t, and V_{ti} are the number of mutual followings author *t* and user *ti* shared, the number of followings author *t* had, and the number of followings *i* had, respectively (a similar "neighborhood overlap" was defined by Onnela et al. 2007). Similarly, we can define and compute overlap indexes of followers $\left(OI_{ti}^{W_1}, OI_{ti}^{W_2}\right)$ by changing *V* to *W* in equation (8). Note that the two numerators in equation (8) are the same: $\overline{V}_{ti}$. The difference between $OI_{ti}^{V_1}$ and $OI_{ti}^{V_2}$ is in the denominators, or in the way by which we scale down $\overline{V}_{ti}$ based on the number of followings *ti* has. Both indexes are in the range [0, 1] because $\overline{V}_{ti} \leq \min\{V_t, V_{ti}\}$. The larger the indexes are, we say the more similar the two sets of followings are. When *t* and *ti* have no mutual followings shared, both indexes equal 0. When *t* and *ti* have exactly the same sets of followings, $OI_{ti}^{V_1}$ =1. When *ti*'s followings represent a subset/superset of *t*'s followings, $OI_{ti}^{V_2}$ =1.

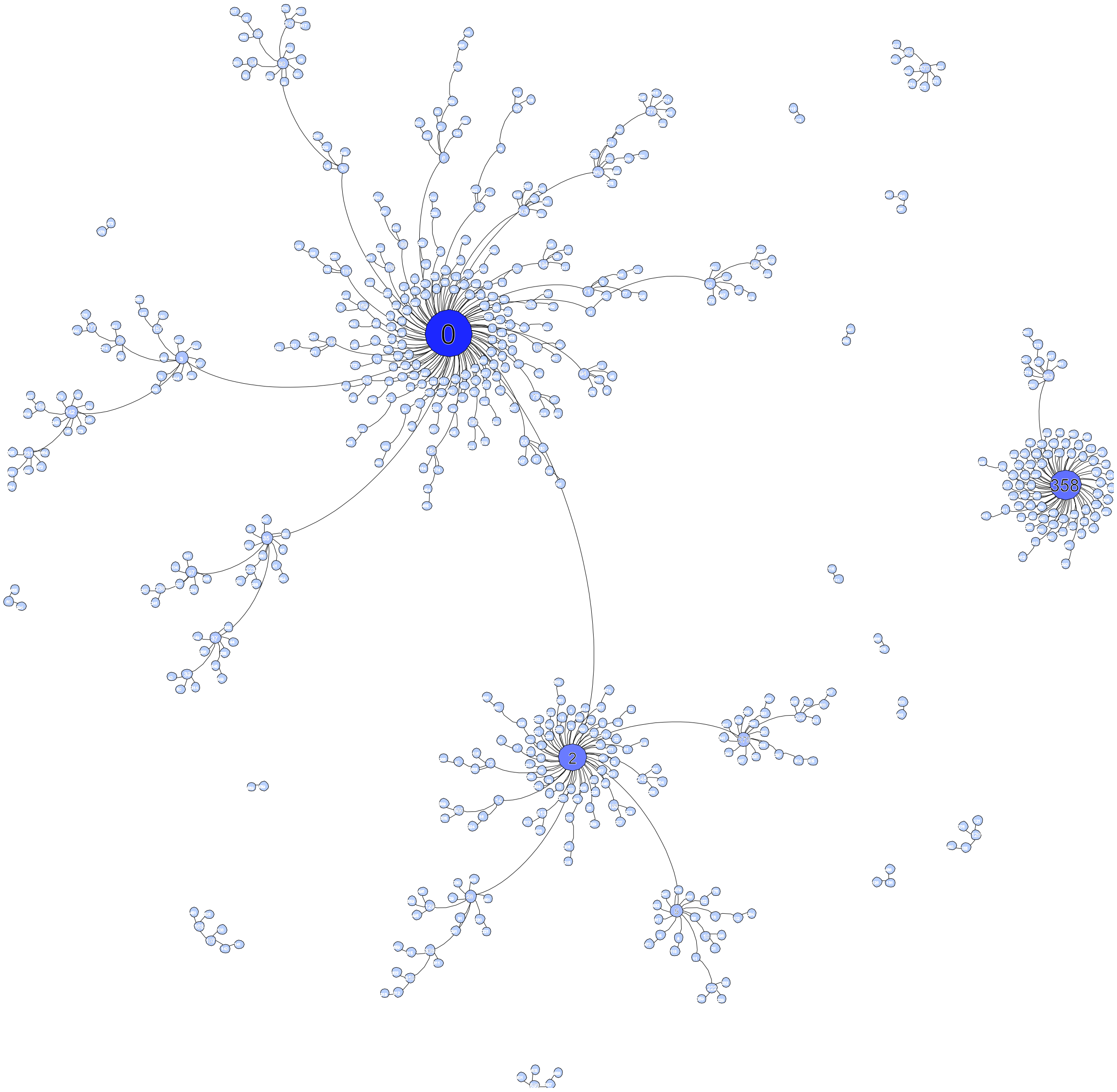
We investigate wether different w_{ti} values lead to significantly different overlap indexes by running a series of ANOVA tests, the results of which are given in Table A1. In all four tests, we control tweet-specific effects. As the regression coefficients in the first row show, we find that a unidirectional relationship w_{ti} = 1) is indeed associated with a smaller overlap in social neighborhoods. The *F* statistics and *p*-values indicate this difference is significant at the 0.1% level, no matter which index we use. Therefore, bidirectional relationships are associated with higher transitivity in social neighborhoods. The results thus support our hypothesis that unidirectional relationships are, on average, weaker than bidirectional ones.

Table A1. Results of Anova Tests

	OI^{V_1}	OI^{V_2}	OI^{W_1}	OI^{W_2}
w_{ti}	-0.042***	-0.069***	-0.034***	-0.064***
F	(2322.21)	(1476.43)	(3837.34)	(2158.65)
p-value	0.00	0.00	0.00	0.00

Appendix B

The Spead of a Single Tweet (t = 1) in Our Sample



Appendix C

Immediate-Follower Retweeters and Other Retweeters

In Table C1, we provide a breakdown of different types of retweeters for each tweet in our sample, including the bidirectional followers of the original author, the unidirectional followers of the original author, those second or higher order retweeters, and other retweeters who are either non-connected or protected. The average ratio of other retweeters is about 38 percent.[1] These other retweeters are most likely users who became exposed to the tweets in our sample after they searched certain keywords because tweets classified as Top Tweets often appear in the top part of the first page of search results if they match the keywords.[2]

Table C1. Number of Immediate-Follower Retweeters and Other Retweeters

t	Immediate Followers		Second and Higher Order Retweeters	Other Retweeters
	Bidirectional	Unidirectional		
1	78	10	545	42
2	7	3	681	787
3	0	11	927	493
4	5	6	423	351
5	2	4	452	1
6	0	8	413	70
7	1	6	399	265
8	2	31	775	341
9	2	6	615	229
10	1	1	292	230
11	1	14	584	46
12	3	15	399	52
13	2	3	264	310
14	7	46	246	117
15	1	9	471	17
16	4	25	435	16
17	3	15	437	110
18	2	13	437	213
19	0	6	363	433
20	2	61	627	58
21	0	27	263	290
22	0	16	275	296
23	1	9	336	319
24	2	10	175	122
25	2	9	256	96
26	1	5	134	280
27	0	23	226	9
28	2	13	350	306
29	70	64	631	87
30	6	1	309	418
31	3	10	215	371
32	1	15	361	281

[1] Note that this ratio is significantly larger than the ratio for randomly selected tweets (Appendix D). For example, among 52 retweets from 200 randomly selected tweets, 32 retweeters are immediate followers of the original authors, 14 are higher order followers, and 6 are from other retweeters.

[2] Indeed, one phenomenon that is consistent with our conjecture is that the proportion of "other retweeters" is significantly higher for tweets with a hashtag (45%) than for tweets without (34%). This is because when people click on a hashtag, Twitter automatically shows the search results containing that particular hashtag.

Table C1. Number of Immediate-Follower Retweeters and Other Retweeters (Continued)

	Immediate Followers		Second and Higher Order	
t	Bidirectional	Unidirectional	Retweeters	Other Retweeters
33	0	14	798	133
34	0	6	113	576
35	2	8	487	14
36	1	10	248	548
37	0	9	230	537
38	8	17	378	85
39	0	30	111	502
40	0	43	261	219
41	0	5	263	487
42	1	5	438	31
43	1	17	293	531
44	3	1	104	12
45	0	36	148	71
46	1	41	113	21
47	0	8	297	748
48	1	8	312	405
49	1	25	669	347
50	2	4	91	19
51	1	11	254	512
52	0	25	373	377
53	0	3	257	419
54	1	12	234	291
55	1	19	136	114
56	0	9	459	703
57	1	3	575	358
58	0	2	382	512
59	6	0	515	726
60	1	3	625	116
61	2	17	227	348
62	12	25	607	16
63	3	13	453	289
64	1	7	922	31
65	0	21	170	593

Appendix D

Random Sample

To investigate the generalizability of our results further, we also collected a random sample of (relatively recent) tweets from the entire Twittersphere (as opposed to TopTweets alone). In this appendix, we show key statistics from this sample. Interested readers can compare them with the ones shown in the main text.

This random sample contains 200 tweets, which were selected from Twitter's (official) public time line in September 2012.[3] We tracked the retweeting activity over a period of two weeks. In terms of the network structure, we collected the IDs of both the followings and the followers of the 200 authors, so that we could identify the unidirectional versus the bidirectional followers.

The 200 authors have 110,672 followers in total. Across the authors, the mean number of followers is 558, the median is 207, the minimum is 0, and the maximum is 13,894 (compared with the first row in Table 2); the mean percentage of proportion of unidirectional followers is 41.9%, the median is 36.4%, the minimum is 0.0%, and the maximum is 100.0% (compared with Figure 7). So, on average, the authors in the TopTweets sample (used in the main text) have a larger proportion of unidirectional followers.

Of the 200 tweets, 20 generated at least 1 retweet for a total of 52 retweets, distributed as follows: 1 tweet generated 16, 1 tweet generated 10, 3 tweets generated 3 each, 2 tweets generated 2 each, and 13 tweets generated 1 each. The authors' immediate followers account for 32 of the 52 retweets. In other words, the retweeting rate among the immediate followers is about 0.03 percent (32/110,672), much lower than that of the TopTweets sample (4.27%), which is not a surprise. *The 32 immediate followers are all unidirectional followers*. This finding is consistent with our key result in the main text.

Several key statistics that depict the distribution of the number of followings and followers are given in Table D1, which can be compared with the third and fourth rows of Table 3.[4] The statistics of the random sample are all larger than those of the TopTweets sample. This finding may be due to the growth of the Twitter network over the more-than-two-year period between July 2010 and September 2012.

Table D1. Descriptive Statistics of the Random Sample

		Mean	*std*	5%	15%	50%	85%	95%
V_{ti}	followings	3,223	16,475	45	120	552	1,984	9,274
W_{ti}	followers	4,527	73,676	11	42	289	1,594	12,300

References

Friedkin, N. 1982. "Information Flow Through Strong and Weak Ties in Intraorganizational Social Networks," *Social Networks* (3), pp. 273-285.

Granovetter, M. 1973. "The Strength of Weak Ties," *The American Journal of Sociology* (78:6), pp. 1360-1380.

Marlow, C., Byron, L., Lento, T., and Rosenn, I. 2009. "Maintained Relationships on Facebook," Overstated (http://overstated.net/2009/03/09/maintained-relationships-on-facebook).

Onnela, J., Saramäki, J., Hyvõnen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. 2007. "Structure and Tie Strengths in Mobile Communication Networks," *Proceedings of the National Academy of Sciences of the U.S.A.* (104:18), pp. 7332-7336.

[3] The public time line is an aggregated stream of all public tweets. We wrote a program to visit the public time line and pick the most recently published tweets. We ran the program every hour to ensure that the tweets were uniformly distributed across the clock hours.

[4] We know the *number* of followings and followers of the 110,672 users, but not their *IDs*.